Continuous-Time Neurodynamic

Nearest Neighbor Classification

Part Two - On the Local Maxima of Sum of Equal-Variance Gaussians

Mehmet Kerem Muezzinoglu

CI Lab. Seminar

November 10, 2004
Problem

Given $\ell$ Gaussians centered at $c_1, \ldots, c_\ell$, all with equal variances, how many local maxima does their sum possess? How far are they located from the centers?

$$E(\cdot): \mathbb{R}^n \to \mathbb{R}, \quad E(x) = \sum_{i=1}^\ell \exp\left(-\gamma \|x - c_i\|_2^2\right)$$
Motivation

This problem has arisen in the RBF-based continuous-time nearest neighbor classifier design - while setting the energy function. (Recall my previous talk)

The goal was to derive a continuous-time dynamical system operating on $[0, 1]^n$ which possesses stable points solely at desired points (prototypes) $p_1, \ldots, p_\ell \in [0, 1]^n$.

We were approaching the problem from an energy function perspective:

$$\dot{x} = \nabla_x E(x)$$

where

$$E(x) = \sum_{i=1}^{\ell} \exp \left( -\gamma \|x - c_i\|_2^2 \right)$$

and $c_i = p_i$. 
Proposition 1

Let $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$ be two Gaussians centered at $a$ and $b$, respectively, both with the same width parameter $\gamma > 0$.

Then, the sum

$$E(x) = \varphi_1(x) + \varphi_2(x)$$

has three extrema if and only if

$$\gamma > \frac{2}{\|a - b\|_2^2}.$$ 

In this case, these points are comprised of one local minimum and two local maxima, all located on the line segment

$$\{x : (1 - \lambda)a + \lambda b, \lambda \in (0, 1)\}$$

The minimum is the midpoint of the line segment, as well as the midpoint of local maxima.
Proof

Definition: A point \( x \in \mathbb{R}^n \) is an extremum of \( E(\cdot) \) iff \( \nabla_x E(x) = 0 \).

Let \( x^* \)

\[
\nabla_x \left( \exp \left( -\gamma \|x^* - a\|_2^2 \right) + \exp \left( -\gamma \|x^* - a\|_2^2 \right) \right) = 0
\]

\[
0 = (x^* - a) \exp \left( -\gamma \|x^* - a\|_2^2 \right) + (x^* - a) \exp \left( -\gamma \|x^* - b\|_2^2 \right) \\
= u(x^* - a) + v(x^* - b)
\]

Then, we have

\[
x^* = \frac{u}{u + v} a + \frac{v}{u + v} b = (1 - \lambda) a + \lambda b.
\]

(2)
Proof (Cont’d)

Substituting (2) in (1), we obtain

\[ 0 = \lambda(a - b) \left( \exp \left( -\gamma \| a - b \|_2^2 \right) \right)^2 \]
\[ + (\lambda - 1)(a - b) \left( \exp \left( -\gamma \| a - b \|_2^2 \right) \right)^{(\lambda - 1)^2} \]
\[ = (a - b) \left[ \lambda \alpha^{\lambda^2} + (\lambda - 1)\alpha^{(\lambda - 1)^2} \right], \tag{3} \]

where \( \alpha = \exp(-\gamma \| a - b \|_2^2) \). (Note: \( 0 < \alpha < 1 \))

Since \( a \) and \( b \) are fixed, we are looking for the solution of

\[ f_\alpha(\lambda) = \lambda \alpha^{\lambda^2} + (\lambda - 1)\alpha^{(\lambda - 1)^2} = 0 \]

to locate the extremum points.
Proof (Cont’d)

Here is the family of curves $f_\alpha(\lambda)$ for several $\alpha$ values.

Note that all curves intersect $\lambda$-axis at $\lambda = 0.5$. So, the midpoint of $a$ and $b$ is an extremum, irrespective of $\alpha$, i.e. $a$ and $b$. 
Proof (Cont’d)

Note from the figure that the curves with a negative slope at the origin intersect the vertical axis at two other points than $\lambda = 0.5$. (and others do not).

$$\frac{df}{d\lambda}(0.5) = \alpha^{\frac{1}{4}}(2 + \ln \alpha) < 0 \iff \alpha < \exp(-2) \iff \gamma > \frac{2}{\|a - b\|^2}$$
Result

The conventional notation of Gaussian is

\[ \exp \left( -\frac{\| x - c \|^2}{2\sigma^2} \right), \]

where \( \sigma \) is called the standard deviation.

By comparing with our notation here, we have

\[ \gamma = \frac{1}{2\sigma^2} \]

“The sum of two Gaussians preserves two local maxima as long as the distance between their centers is greater than \( 2\sigma \).

Otherwise, there occurs a unique extremum at the midpoint, which is a maximum.”
Result (cont’d)

An individual Gaussian has a unique maximum at its center.

Summing up two Gaussians (centered at distinct points) pulls the individual maxima towards each other.

This can be generalized as follows: A new Gaussian added to the sum pulls all maxima (around other centers) towards its own center.

The amount of displacement is dependent upon the distances between the centers, as well as the width parameter $\gamma$. Smaller $\gamma$ results in larger displacements.
Question

How can we ensure that exactly one local maximum occurs in the vicinity of each center when we sum $\ell$ Gaussians up?

For $\ell = 2$, the answer is straightforward due to the first result above:

Let’s denote our tolerance for each center as $\epsilon > 0$. To have a solution of (1) within this tolerance, we allow only

$$0 < \lambda \leq \epsilon / \| a - b \|_2.$$

Then, we set $f(\epsilon / \| a - b \|_2) = 0$ and solve for $\gamma$. This, or any greater, $\gamma$ satisfies our constraint.
Definitions

- Convex hull of $m$ points $p_1, \ldots, p_m$ in $\mathbb{R}^n$ is

$$\mathcal{H} = \{x \in \mathbb{R}^n : x = \alpha_1 p_1 + \alpha_2 p_2 + \ldots + \alpha_m p_m, \alpha_i \geq 0, \sum_{i=1}^m \alpha_i = 1\}$$

- Ball (or sphere as sometimes called) with radius $\epsilon$ centered at a point $p$ is

$$B_\epsilon(p) = \{x \in \mathbb{R}^n : \|x - p\|_2 < \epsilon\}$$
Proposition 2

Consider $\ell$ Gaussians $\varphi_1(\cdot), \varphi_2(\cdot), \ldots, \varphi_\ell(\cdot)$ centered at $c_1, c_2, \ldots, c_\ell$ with the same width parameter $\gamma$ such that

$$\gamma > \frac{2}{\min_{i \neq j} \|c_i - c_j\|^2_2},$$

i.e. $\varphi_i(\cdot) + \varphi_j(\cdot)$ preserves two local maxima for all $i, j \in \{1, \ldots, \ell\}$.

Let $\lambda^*_{ij}$ be the (unique) solution of

$$(\lambda - 1) \left( \exp \left( -\gamma \|c_i - c_j\|^2_2 \right) \right)^{(\lambda-1)^2} + \lambda \left( \exp \left( -\gamma \|c_i - c_j\|^2_2 \right) \right) \lambda^2 = 0$$

within $\lambda \in (0, 0.5)$ for all $i, j \in \{1, \ldots, \ell\}$.

Then, for each $i \in \{1, \ldots, \ell\}$, there exists a unique local maximum within the ball $B_{\epsilon_i}(c^i)$ where

$$\epsilon_i = \left\| \sum_{j \neq i} \lambda^*_{ij} \|c_i - c_j\|_2 \right\|_2.$$
Proof
Later in the appendix...
Corruption of Prototypes is Unavoidable but Reducable

In any case, adding two or more Gaussians results in local maxima shifted from the desired prototype patterns.

According to the results above we see that, as the width parameter increases, local maxima of

\[ E(x) = \sum_{i=1}^{\ell} \exp \left( -\gamma \| x - c_i \|_2^2 \right) \]

move towards the centers. So, what we can do to improve the representation is to keep this corruption as small as possible by choosing \( \gamma \) sufficiently large.

Theoretically, it is wise to pick the design parameter \( \gamma \) very large.
Experimental Verification

100 uniform i.i.d. instances of NN classification has been generated randomly. (Prototypes and distorted pattern)

For each problem instance, the gradient system to maximize $E(\cdot)$ has been designed for each of the 6 width parameters: $\gamma_{\text{inf}} + .1$, $\gamma_{\text{inf}} + .5$, $\gamma_{\text{inf}} + 1$, $\gamma_{\text{inf}} + 10$, $\gamma_{\text{inf}} + 20$, $\gamma_{\text{inf}} + 30$, where $\gamma_{\text{inf}}$ denotes the absolute lower bound on $\gamma$ for the particular prototype distribution.

<table>
<thead>
<tr>
<th>$\gamma_{\text{inf}}$</th>
<th>$\gamma_{\text{inf}}^{0.1}$</th>
<th>$\gamma_{\text{inf}}^{0.5}$</th>
<th>$\gamma_{\text{inf}}^{1}$</th>
<th>$\gamma_{\text{inf}}^{10}$</th>
<th>$\gamma_{\text{inf}}^{20}$</th>
<th>$\gamma_{\text{inf}}^{30}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC%</td>
<td>20</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$$MC\% = \frac{\# \text{ misclassified instances}}{\# \text{ all instances}} \cdot 100$$
Does a Sufficiently Large $\gamma$ Really Solve the All Problems?

This solves the representation problem. (Creates a more similar fixed point to the associated prototype).

However, it gives rise to a major puzzle on gradient system’s performance.
Large $\gamma$ Slows down the Convergence

The gradient is almost zero all over the state space except in neighborhoods of the centers. So, for initial conditions excluded by these neighborhoods, the system performs very slow.

As $\gamma$ increases, these neighborhoods shrink.

\[
\begin{array}{c|c|c|c|c|c|c}
\gamma_{\text{inf}} & \gamma_{\text{inf}}^{0.1} & \gamma_{\text{inf}}^{0.5} & \gamma_{\text{inf}}^1 & \gamma_{\text{inf}}^{10} & \gamma_{\text{inf}}^{20} & \gamma_{\text{inf}}^{30} \\
\hline
\text{N/C}\% & 0 & 0 & 9 & 18 & 37 & 46 \\
\end{array}
\]

\[
N/C\% = \frac{\# \text{ nonconvergent instances}}{\# \text{ all instances}} \cdot 100
\]

Here we mean by nonconvergent instant, an instant for which the continuous-time gradient system cannot produce any decision within 250 seconds.
An extreme instant where the system cannot perform reasonably fast for any admissible $\gamma$:

$$\nabla E(x_0) = 10^{-6} \cdot (.2, -.3)$$

Too small in magnitude!
Even for $\gamma = 25$

$$||c_1 - c_2||_2^2 = .08 \Rightarrow \gamma > 25$$
Solution: Varying $\gamma$ along Dynamics

In problem instances where we cannot find a solution within the trade-off between system velocity and accuracy (small $\gamma$ versus large $\gamma$), we have to give up fixed $\gamma$ assumption.

There are two ways to vary $\gamma$ in order to achieve optimum performance:

- Assume $\gamma = \gamma(x)$ is a nonlinear function of the state variables.
- Assume $\gamma = \gamma(t)$ has explicit dependence on time.

This is actually the choice of making the system time-varying or more nonlinear.

We adopt the second one at this point.
Thank You!

To be continued...